JITET (Jurnal Informatika dan Teknik Elektro Terapan)
Vol. 13 No. 1, pISSN: 2303-0577 eISSN: 2830-7062 http://dx.doi.org/10.23960/jitet.v13i1.5864

OPTIMASI ALGORITMA K-NEAREST (KNN) NEIGHBORS PADA PREDIKSI RISIKO PENYAKIT KARDIOVASKULAR

Peri Hidayat[1*], Rudi Kurniawan[2], Yudhistira Arie Wijaya[3], Tati Suprapti[4].

[1,2,3,4]STMIK IKMI Cirebon; Jl.Perjuangan No.10B, Majasem, Cirebon, Jawa Barat 45135; Telp. (0231)490480

Received: 28 Desember 2024
Accepted: 14 Januari 2025
Published: 20 Januari 2025

Keywords:
K-Nearest Neighbors;
Penyakit kardiovaskular;
Knowledge Discovery in Databases;
Akurasi;
Nilai K;

Corespondent Email:
perihidayat2000@gmail.com

Abstrak. Penyakit kardiovaskular merupakan penyebab utama kematian di dunia, dipengaruhi oleh berbagai faktor risiko yang kompleks. Deteksi dini sangat penting untuk mencegah komplikasi serius. Penelitian ini bertujuan mengembangkan model prediksi risiko penyakit kardiovaskular menggunakan algoritma *K-Nearest Neighbors* (K-NN) dan menentukan nilai K optimal untuk meningkatkan akurasi prediksi. Metodologi yang digunakan adalah *Knowledge Discovery in Databases* (KDD), mencakup pemilihan data, pembersihan data, transformasi data, pemilihan atribut, evaluasi, dan validasi model. Dataset yang digunakan terdiri dari variabel medis seperti usia, berat badan, tekanan darah, kadar kolesterol, dan riwayat medis lainnya. Data dibagi dengan rasio 70:30 dan 80:20 untuk mengevaluasi performa model pada pembagian data yang berbeda. Hasil menunjukkan bahwa nilai K = 40 memberikan akurasi terbaik sebesar 71,00% pada rasio 70:30, sedangkan nilai K = 25 menghasilkan akurasi 71,16% pada rasio 80:20. Kesimpulan penelitian ini adalah algoritma K-NN mampu memprediksi risiko penyakit kardiovaskular dengan baik, bergantung pada pemilihan nilai K dan rasio pembagian data yang optimal. Penelitian ini berkontribusi dalam pengembangan model prediksi risiko penyakit kardiovaskular dan menjadi referensi untuk diagnosis dini di masa depan.

Abstract. *Cardiovascular disease is a leading cause of death worldwide, influenced by various complex risk factors. Early detection is crucial to prevent severe complications. This study aims to develop a predictive model for cardiovascular disease risk using the K-Nearest Neighbors (K-NN) algorithm and determine the optimal K value to enhance prediction accuracy. The methodology applied is Knowledge Discovery in Databases (KDD), covering data selection, cleaning, transformation, attribute selection, evaluation, and model validation. The dataset includes medical variables such as age, weight, blood pressure, cholesterol levels, and medical history. Data were split into two ratios, 70:30 and 80:20, to assess model performance under different data partitions. Results showed that K = 40 achieved the best accuracy of 71.00% in the 70:30 ratio, while K = 25 yielded the highest accuracy of 71.16% in the 80:20 ratio. This study concludes that the K-NN algorithm effectively predicts cardiovascular disease risk, depending on the appropriate choice of K value and data partition ratio. The research significantly contributes to the advancement of predictive models for cardiovascular disease risk and serves as a reference for early diagnosis in future studies.*

1. PENDAHULUAN

Kemajuan teknologi informasi dan ilmu komputer telah membawa pengaruh besar di berbagai bidang kehidupan, seperti kesehatan, pendidikan, dan bisnis. Dalam bidang kesehatan, penggunaan teknologi informasi berbasis data menjadi krusial untuk merumuskan rekomendasi kebijakan yang lebih efektif dan efisien [1]. Di era digital saat ini, pengelolaan data yang semakin melimpah menjadi hal yang sangat krusial. Berbagai studi mengungkapkan bahwa penggunaan teknologi dan sistem informasi yang tepat mampu meningkatkan efisiensi dalam pengelolaan data. [2]. Penyakit kardiovaskular, seperti penyakit jantung koroner dan stroke, merupakan penyebab utama kematian di dunia, sehingga metode prediksi yang akurat sangat penting dalam penelitian ini. Algoritma K-Nearest Neighbor (KNN) telah terbukti sangat fleksibel dalam menangani masalah klasifikasi yang kompleks. Penelitian ini berfokus pada penerapan algoritma K-Nearest Neighbors (KNN) untuk mengklasifikasikan penyakit kardiovaskular [3]. Oleh karena itu, pengembangan metode klasifikasi yang akurat dan efektif dalam diagnosis penyakit kardiovaskular sangat penting. Deteksi dini dan penanganan yang tepat bagi penderita penyakit kardiovaskular tidak boleh diabaikan. Ketidakstabilan kadar gula darah dapat berdampak serius pada organ tubuh, meningkatkan risiko penyakit kardiovaskular, merusak ginjal, dan menyebabkan berbagai komplikasi lainnya. Oleh karena itu, sangat diperlukan pengembangan metode klasifikasi yang akurat dan efektif dalam mendiagnosis penyakit kardiovaskular [3].

Tantangan dalam memprediksi penyakit kardiovaskular terletak pada kebutuhan akan metode yang dapat menghasilkan hasil yang akurat dengan waktu pengerjaan yang efisien [4]. Untuk meningkatkan akurasi model tanpa mengurangi kecepatan pemrosesan, berbagai penelitian telah dilakukan dengan menerapkan pendekatan algoritma yang berbeda. [5]. Dalam pengembangan algoritma pembelajaran mesin, khususnya K-Nearest Neighbors (KNN), terdapat tantangan besar terkait dengan optimalisasi parameter serta pengolahan data yang efektif. Meskipun berbagai penelitian telah mengungkapkan potensi KNN dalam beragam aplikasi, isu ini tetap menjadi perhatian utama untuk pengembangan di masa depan [6]. Oleh sebab itu, diperlukan pendekatan yang tidak hanya mampu meningkatkan akurasi, tetapi juga menjamin keandalan dan efisiensi sistem. Beragam studi telah dilakukan untuk mengeksplorasi berbagai metode yang dapat mendukung peningkatan akurasi prediksi [7]. Selain itu, data yang tidak seimbang dan adanya nilai yang hilang sering kali menjadi kendala dalam proses pengolahan data secara maksimal [8]. Oleh karena itu, pemilihan algoritma serta teknik optimasi yang sesuai menjadi faktor krusial untuk menghasilkan prediksi risiko penyakit yang lebih akurat dan aplikatif dalam praktik. Dalam hal ini, sejumlah penelitian terkini mengungkapkan kemajuan yang signifikan dalam penerapan algoritma optimasi guna meningkatkan akurasi prediksi di bidang kesehatan. [9].

Penelitian terdahulu telah mengeksplorasi beragam teknik pembelajaran mesin untuk klasifikasi penyakit kardiovaskular. Salah satu studi yang memiliki keterkaitan erat adalah penelitian yang dilakukan oleh [10] yang mengaplikasikan teknologi machine learning untuk klasifikasi jenis hipertensi berdasarkan fitur pribadi, menunjukkan peningkatan akurasi yang signifikan dalam klasifikasi penyakit kardiovaskular. Dalam upaya untuk membandingkan kinerja beberapa algoritma dalam klasifikasi penyakit jantung, terdapat sejumlah penelitian yang menunjukkan bahwa algoritma K-Nearest Neighbors (K-NN) memberikan akurasi yang kompetitif, meskipun peningkatan efisiensi masih diperlukan. Salah satu penelitian oleh [11] membandingkan algoritma K-NN dengan Random Forest dalam konteks penyakit gagal jantung, dan menemukan bahwa K-NN dapat memberikan hasil yang memuaskan dalam klasifikasi. Penelitian ini juga menekankan pentingnya pemilihan algoritma yang tepat untuk meningkatkan akurasi diagnosis penyakit jantung. Dalam konteks ini, beberapa penelitian telah menunjukkan bahwa penggunaan algoritma yang sesuai dapat secara signifikan mempengaruhi hasil klasifikasi. Misalnya, penelitian oleh [12] menunjukkan bahwa algoritma Naïve Bayes dapat digunakan untuk menganalisis data penyakit jantung koroner,

dengan hasil yang menunjukkan performa yang baik dalam klasifikasi. Hasil penelitian tersebut memberikan peluang untuk penelitian lebih lanjut yang berfokus pada pengembangan teknik yang dapat mengoptimalkan akurasi tanpa mengurangi efisiensi. Dalam konteks ini, beberapa penelitian terkini menunjukkan bahwa pendekatan berbasis teknologi informasi dapat meningkatkan efisiensi dan akurasi dalam berbagai bidang [13].

Penelitian ini bertujuan untuk meningkatkan model prediksi risiko penyakit kardiovaskular, dengan mengandalkan algoritma K-Nearest Neighbor (KNN) yang telah dioptimalkan sebagai salah satu pendekatan yang potensial [14]. Kontribusi utama dari penelitian ini terletak pada pengembangan pendekatan yang bertujuan untuk mengurangi kesalahan klasifikasi dalam sistem peringatan dini penyakit kardiovaskular [15]. Dengan demikian, penelitian ini diharapkan dapat mengatasi kesenjangan dalam literatur yang ada sekaligus memberikan kontribusi yang berarti di bidang informatika kesehatan. Selain itu, penelitian ini bertujuan membantu dokter dalam mendiagnosis dan mengelola risiko penyakit secara lebih efektif. Dalam hal ini, sejumlah studi terbaru menyoroti pentingnya penerapan sistem informasi dalam sektor kesehatan [16].

Dalam penelitian ini, teknik Knowledge Discovery in Database (KDD) diterapkan sebagai metode utama. KDD mencakup beberapa langkah penting, yaitu pemilihan data, praproses data, transformasi, pemodelan, dan evaluasi. [17]. Pada tahap pemodelan, algoritma K-Nearest Neighbor (KNN) diterapkan dan dievaluasi untuk menentukan nilai K yang optimal guna mencapai akurasi tertinggi. Pendekatan ini diharapkan dapat mengatasi tantangan pengolahan data yang kompleks serta menghasilkan model prediksi yang akurat dan efisien. Berbagai penelitian telah menunjukkan penerapan KNN dalam berbagai bidang, seperti klasifikasi penyakit, analisis sentimen, dan sistem rekomendasi [18].

2. TINJAUAN PUSTAKA

2.1 Classification

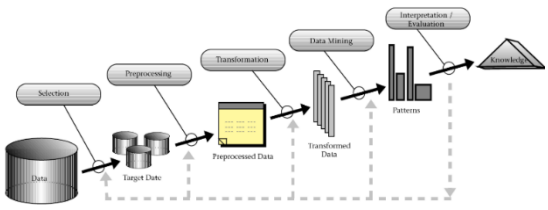
Klasifikasi adalah proses pengelompokan benda atau hal-hal lainnya berdasarkan ciri-ciri tertentu [19]. klasifikasi membagi data ke dalam kelompok-kelompok berdasarkan batasan yang telah ditentukan [20].

2.2 Algoritma K-Nearest Neighbors

Algoritma K-Nearest Neighbors adalah K-Nearest Neighbor (KNN) adalah salah satu metode klasifikasi yang memanfaatkan pola-pola data yang ada dalam dataset untuk mengklasifikasi kategori atau kelas dari suatu sampel yang belum diketahui. Pengklasifikasi K-Nearest Neighbor (KNN) telah membuktikan fleksibilitas yang tinggi dalam permasalahan klasifikasi yang kompleks. [3].

3. METODE PENELITIAN

Penelitian ini menerapkan metode kuantitatif dengan pendekatan eksperimen. Metode kuantitatif bertujuan untuk mengembangkan model matematis, teori, dan hipotesis yang berkaitan dengan suatu fenomena, guna menentukan hubungan antar variabel dalam sebuah populasi. Proses pengukuran berfungsi menghubungkan pengamatan empiris dengan analisis matematis dari hubungan kuantitatif yang ada [21]. Dalam penelitian ini menggunakan metode KDD untuk menganalisa data. Tahapan metode KDD seperti tampak pada Gambar 1.1.



Gambar 1. 1 Tahapan Proses KDD

3.1 Data Understanding

Data understanding merupakan fase penting dalam metodologi ilmu data dan perkembangan kecerdasan buatan, dengan tujuan memperoleh wawasan awal tentang data yang diperlukan untuk mengatasi tantangan bisnis tertentu.

3.2 Selection

Proses pemilihan data berkaitan dengan pengurangan volume data yang digunakan dalam proses penambangan, sementara secara bersamaan memastikan bahwa data asli terwakili secara memadai. Data yang dipilih untuk dimasukkan ke dalam prosedur

penambangan data disimpan dalam file yang berbeda, terpisah dari database produksi.

3.3 Preprocessing

Preprocessing merupakan prosedur sistematis untuk membersihkan kumpulan data atau mengidentifikasi dan memperbaiki (atau menghilangkan) entri yang dikompromikan atau salah dari kumpulan catatan, tabel, atau database. Ini mencakup membedakan data yang tidak lengkap, salah, atau asing dan modifikasi atau penghapusan data yang terkontaminasi atau cacat.

3.4 Transformation

Transformation data adalah proses mengubah data dari satu format ke format lainnya. Transformasi data mencakup konversi sistematis data dari satu format ke format lainnya. Bentuk dominan transformasi data melibatkan transisi data yang tidak dimurnikan ke dalam format yang dipoles dan fungsional, mengubah tipe data, menghilangkan data berlebihan, dan menambah data untuk meningkatkan kemanjuran organisasi. Sepanjang prosedur transformasi data, analis memastikan kerangka kerja, mengeksekusi pemetaan data untuk mengambil data dari sumber awal, melakukan transformasi, dan akhirnya mengarsipkan data dalam database yang ditunjuk.

3.5 Data Mining

Data Mining mencakup penerapan metodologi dan teknik khusus untuk mengidentifikasi pola dan mengekstrak informasi penting dari kumpulan data yang dikuratori. Keragaman teknik, metodologi, dan algoritma penambangan data sangat luas. Pemilihan metode atau algoritma yang tepat sebagian besar bergantung pada tujuan spesifik dan kerangka kerja Penemuan Pengetahuan dalam Database (KDD) menyeluruh. Tujuan utamanya adalah untuk menganalisis kumpulan data substansial untuk membedakan tren, pola, dan keterkaitan yang memfasilitasi pengambilan keputusan dan perencanaan strategis yang diinformasikan. Dalam penelitian ini, teknik algoritma Naive Bayes dan K-Nearest Neighbors digunakan untuk proses ekstraksi informasi.

3.6 Evaluation and Interpretation

Interpretasi/evaluasi merupakan analisis atau sintesis hasil yang berasal dari proses penambangan data. Kesimpulan akhir ditetapkan dengan menggabungkan beberapa hipotesis yang dihasilkan melalui metodologi penambangan data.

4. HASIL DAN PEMBAHASAN

Tahapan Penelitian ini mengacu pada metodologi yang diterapkan yaitu metode KDD (Knowledge Discovery in Database).

4.1 Data Understanding

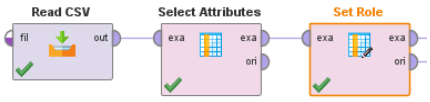
Data yang digunakan dalam penelitian ini diambil dari situs Kaggle dalam format file CSV dengan nama "heart_data.csv", yang merupakan dataset publik yang berisi informasi terkait diagnosis penyakit kardiovaskular. Dataset ini terdiri dari 70.000 baris data dan mencakup 14 atribut utama yang menggambarkan berbagai faktor risiko serta karakteristik pasien, seperti usia, jenis kelamin, tekanan darah, kadar kolesterol, dan riwayat kesehatan lainnya.

Tabel 1. 1 Atribut Data dari file Data heart_data

No	Atribut	Tipe Data	Keterangan
1.	*index*	*Integer*	Nomor urut data sebagai indeks, tanpa satuan.
2.	*id*	*Integer*	Identifikasi unik untuk setiap entri dalam dataset.
3.	*age*	*Integer*	Usia subjek dalam hitungan hari
4.	*gender*	*Integer*	Jenis kelamin: 1 untuk pria, 2 untuk wanita.
5.	*height*	*Integer*	Tinggi badan subjek dalam satuan cm (*centimeter*).
6.	*weight*	*Real*	Berat badan subjek dalam satuan kg (kilogram).
7.	*ap_hi*	*Integer*	Tekanan darah sistolik dalam satuan mmHg (milimeter merkuri).
8.	*ap_lo*	*Integer*	Tekanan darah diastolik dalam satuan mmHg (milimeter merkuri).
9.	*cholesterol*	*Integer*	Kadar kolesterol: 1 = normal, 2 = di atas normal, 3 = jauh di atas normal.
10	*gluc*	*Integer*	Kadar glukosa: 1 = normal, 2 = di atas normal, 3 = jauh di atas normal.
11.	*smoke*	*Integer*	Apakah subjek merokok: 0 = tidak, 1 = ya.
12.	*alco*	*Integer*	Apakah subjek mengonsumsi alkohol: 0 = tidak, 1 = ya.
13.	*active*	*Integer*	Apakah subjek aktif secara fisik: 0 = tidak, 1 = ya.
14.	*cardio*	*Integer*	Status kardiovaskular: 0 = tidak memiliki penyakit kardiovaskular, 1 = memiliki penyakit.

4.2 Selection

Pemilihan dataset dilakukan sebelum tahap eksplorasi data dalam KDD. Langkah pertama yang dilakukan adalah mengimpor dataset "heart_data" menggunakan operator Read CSV. Adapun model di Selection seperti pada Gambar .



Gambar 4. 1 Model Data Selection

Kemudian hasil dari analisis model dari tiap tiap operator akan mendapatkan hasil seperti pada Tabel.

Tabel 4. 1 Hasil Operator Read Csv

index	id	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio
0	0	18393	2	168	62.0	110	80	1	1	0	0	1	0
1	1	20228	1	156	85.0	140	90	3	1	0	0	1	1
2	2	18857	1	165	64.0	130	70	3	1	0	0	0	1
3	3	17623	2	169	82.0	150	100	1	1	0	0	1	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...
69996	99995	22601	1	158	126.0	140	90	2	2	0	0	1	1
69997	99996	19066	2	183	105.0	180	90	3	1	0	1	0	1
69998	99998	22431	1	163	72.0	135	80	1	2	0	0	0	1
69999	99999	20540	1	170	72.0	120	80	2	1	0	0	1	0

Langkah selanjutnya dalam penelitian ini adalah menggunakan operator Select Attributes, yang berfungsi untuk memilih atau menentukan atribut-atribut yang akan digunakan dalam analisis prediksi risiko penyakit kardiovaskular, Hal ini ditunjukkan pada Tabel 4.2.

Tabel 4. 2 Hasil Operator Select Atribut

index	id	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio
0	0	18393	2	168	62.0	110	80	1	1	0	0	1	0
1	1	20228	1	156	85.0	140	90	3	1	0	0	1	1
2	2	18857	1	165	64.0	130	70	3	1	0	0	0	1
3	3	17623	2	169	82.0	150	100	1	1	0	0	1	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...
69996	99995	22601	1	158	126.0	140	90	2	2	0	0	1	1
69997	99996	19066	2	183	105.0	180	90	3	1	0	1	0	1
69998	99998	22431	1	163	72.0	135	80	1	2	0	0	0	1
69999	99999	20540	1	170	72.0	120	80	2	1	0	0	1	0

pada pengguna operator *Set Role* menunjukkan bahwa atribut dalam dataset telah berhasil diberi peran sesuai dengan fungsinya dan terdapat 2 spesial atribut dan 12 regular atribut. Atribut-atribut seperti *id, index age, gender, height, weight, ap_hi, ap_lo, cholesterol, gluc, smoke, alco,* dan *active* diperlakukan sebagai regular, yang berarti atribut-atribut ini berfungsi sebagai fitur input untuk memprediksi status kardiovaskular. Sementara itu, atribut *cardio* yang menunjukkan status kardiovaskular (0 untuk tidak memiliki penyakit dan 1 untuk memiliki penyakit) diperlakukan sebagai label, yang menjadi target atau hasil yang ingin diprediksi oleh model.

Tabel 4. 3 Hasil Operator Set Role

id	cardio	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active
0	0	18393	2	168	62.0	110	80	1	1	0	0	1
1	1	20228	1	156	85.0	140	90	3	1	0	0	1
2	1	18857	1	165	64.0	130	70	3	1	0	0	0
3	1	17623	2	169	82.0	150	100	1	1	0	0	1
...	...	...	...	...	...	...	...	...	...	...	...	...
99995	1	22601	1	158	126.0	140	90	2	2	0	0	1
99996	1	19066	2	183	105.0	180	90	3	1	0	1	0
99998	1	22431	1	163	72.0	135	80	1	2	0	0	0
99999	0	20540	1	170	72.0	120	80	2	1	0	0	1

4.3 Preprocessing

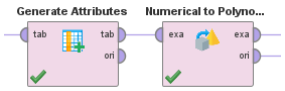
Langkah *preprocessing* data dalam penelitian ini melibatkan proses pembersihan data untuk mengatasi nilai yang hilang atau tidak konsisten. Pertama, dilakukan evaluasi untuk memastikan tidak ada nilai yang hilang atau tidak valid dalam dataset, dan didalam ekperimen kali ini tidak ditemukan data yang missing seperti tampak pada Tabel

Tabel 4. 4 Hasil Preprocessing

No	Attribut	Type Atribut	Missing Values
1.	*cardi*	*Integer*	0
2.	*cholesterol*	*Integer*	0
3.	*weight*	*Integer*	0
4.	*age_in_years*	*Integer*	0
5.	*age*	*Integer*	0
6.	*gender*	*Integer*	0
7.	*height*	*Integer*	0
8.	*weight*	*Integer*	0
9.	*ap_hi*	*Integer*	0
10.	*ap_lo*	*Integer*	0
11.	*gluc*	*Integer*	0
12.	*smoke*	*Integer*	0
13.	*alco*	*Integer*	0
14.	*active*	*Integer*	0

4.4 Transformation

Transformasi ini penting untuk memastikan bahwa data yang telah diproses sebelumnya dapat diolah dengan baik dalam *RapidMiner AI Studio 2024*. Dengan melakukan transformasi yang tepat, seperti mengonversi tipe data atau memilih atribut yang relevan, data akan siap untuk diterapkan dalam model *K-NN*. Adapun Modelnya seperti pada Gambar 4.2.



Gambar 4. 2 Model Transformation

Dalam penelitian ini, *Operator Generate Attribute* digunakan di *RapidMiner* untuk mengubah atribut *age* dari satuan hari menjadi tahun.

Adapun hasilnya seperti pada Tabel 4.5.

Tabel 4. 5 Hasil Operator Generate Atribut

cardio	cholesterol	age	gender	height	weight	ap_hi	ap_lo	gluc	smoke	alco	active	age_in_years
0	1	18393	2	168	62.0	110	80	1	0	0	1	50
1	3	20228	1	156	85.0	140	90	1	0	0	1	55
1	3	18857	1	165	64.0	130	70	1	0	0	0	52
...	...	...	...	...	...	...	...	...	...	...	...	...
1	2	22601	1	158	126.0	140	90	2	0	0	1	62
1	3	19066	2	183	105.0	180	90	1	0	1	0	52
1	1	22431	1	163	72.0	135	80	2	0	0	0	61
0	2	20540	1	170	72.0	120	80	1	0	0	1	56

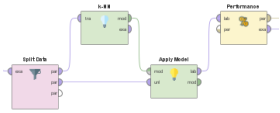
Selanjutnya, operator *Numerical to Polynomial* digunakan untuk mengubah atribut numerik, seperti *age_in_years*, *cholesterol*, dan *weight*, menjadi data kategorikal (nominal) dengan membagi nilai-nilai numerik tersebut ke dalam interval atau kategori tertentu. Adapun hasilnya seperti pada Tabel 4.6

Tabel 4. 6 Hasil Operator Numerical to Polinominal

cardio	cholesterol	weight	age_in_years	age	gender	height	ap_hi	ap_lo	gluc	smoke	alco	active
0	1	62.0	50	18393	2	168	110	80	1	0	0	1
1	3	85.0	55	20228	1	156	140	90	1	0	0	1
1	3	64.0	52	18857	1	165	130	70	1	0	0	0
1	1	82.0	48	17623	2	169	150	100	1	0	0	1
...	...	...	...	...	...	...	...	...	...	...	...	...
1	2	126.0	62	22601	1	158	140	90	2	0	0	1
1	3	105.0	52	19066	2	183	180	90	1	0	1	0
1	1	72.0	61	22431	1	163	135	80	2	0	0	0
0	2	72.0	56	20540	1	170	120	80	1	0	0	1

4.5 Data Mining

Data mining yaitu suatu proses untuk mencari informasi yang menarik dari data yang disimpan dalam jumlah banyak dengan menerapkan teknik atau metode tertentu. Penggunaan *split data* dalam penelitian ini bertujuan untuk membagi dataset menjadi dua bagian, yaitu data training dan data testing. hasil penelitian dan menghindari bias dalam pembagian data. Adapun modelnya seperti pada Gambar 4.3.



Gambar 4. 3 Model Data Mining

Adapun hasil dari split data seperti pada Tabel 4.15.

Tabel 4. 7 Hasil Operator Split Data

cardio	cholesterol	weight	age_in_years	age	gender	height	ap_hi	ap_lo	gluc	smoke	alco	active
0	1	62.0	50	18393	2	168	110	80	1	0	0	1
1	3	85.0	55	20228	1	156	140	90	1	0	0	1
1	3	64.0	52	18857	1	165	130	70	1	0	0	0
1	1	82.0	48	17623	2	169	150	100	1	0	0	1
...	...	...	...	...	...	...	...	...	...	...	...	...
1	2	126.0	62	22601	1	158	140	90	2	0	0	1
1	3	105.0	52	19066	2	183	180	90	1	0	1	0
1	1	72.0	61	22431	1	163	135	80	2	0	0	0
0	2	72.0	56	20540	1	170	120	80	1	0	0	1

Selanjutnya Operator *K-Nearest Neighbors (K-NN)* dalam penelitian ini digunakan untuk membangun model prediksi risiko penyakit kardiovaskular dengan mengklasifikasikan individu berdasarkan kedekatan atribut-atribut yang relevan, seperti *age_in_years*, *cholesterol*, *weight*, dan *cardio*. Adapun parameter yang digunakan dalam Operator K-NN seperti pada Tabel 4.8.

Tabel 4. 8 Parameter K-NN

No.	Parameter	Isi
1.	K	2 - 50
2.	*Weighted Vote*	✓
3.	*Measure Type*	*MixedMeasure*
4.	*Mixed Measure*	*MixedEuclideanDistance*

Terdapat beberapa parameter yang digunakan untuk mengkonfigurasi algoritma *K-Nearest Neighbors (K-NN)* dalam penelitian ini. Parameter K menentukan jumlah tetangga terdekat yang akan dipertimbangkan untuk klasifikasi, dengan rentang nilai antara 2 hingga 50, yang berfungsi untuk menyeimbangkan akurasi prediksi dengan kompleksitas model. Parameter *Weighted Vote* diaktifkan untuk

memperioritaskan data yang lebih relevan, artinya bahwa dari setiap tetangga akan diberikan bobot berdasarkan jaraknya, tetangga yang lebih dekat akan memiliki pengaruh lebih besar terhadap prediksi. Parameter *Measure Type* diset ke *MixedMeasure*, yang memungkinkan penggunaan berbagai tipe metrik jarak, sedangkan *Mixed Euclidean Distance* adalah jenis metrik yang digunakan, yang mengkombinasikan jarak *Euclidean* untuk atribut numerik dan kategorikal.

Selanjutnya Operator *Apply Model* digunakan dalam penelitian ini untuk menerapkan model yang telah dilatih terhadap data uji atau data baru guna memperoleh hasil prediksi. Seperti pada Tabel 4.9.

Tabel 4. 9 Hasil Operator Apply Model

cardio	Prediction(cardio)	Confidence(0)	Confidence(1)	cholesterol	cholesterol	weight	age_in_years	age	gender	height	ap_hi	ap_lo	gluc	smoke	alco	active
0	1	0.040	0.960	1	1	62.0	50	18393	2	168	110	80	1	0	0	1
1	0	0.761	0.239	1	3	85.0	55	20228	1	156	140	90	1	0	0	1
1	0	0.799	0.201	3	3	64.0	52	18857	1	165	130	70	1	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1	1	0.479	0.520	2	2	126.0	62	22601	1	158	140	90	2	0	0	1
1	0	0.680	0.319	1	3	105.0	52	19066	2	183	180	90	1	0	1	0
1	0	0.600	0.340	1	1	72.0	61	22431	1	163	135	80	2	0	0	0
0	0	0.540	0.460	2	2	72.0	56	20540	1	170	120	80	1	0	0	1

Setelah hasil dari *apply model* selanjutnya operator *Performance* digunakan untuk mengevaluasi akurasi model *K-Nearest Neighbors (K-NN)*. *Parameter Main Criterion*: *first* menunjukkan bahwa akurasi dipilih sebagai metrik utama untuk menilai kinerja model. Dengan memilih Accuracy, penelitian ini bertujuan untuk mengukur seberapa tepat model memprediksi risiko penyakit kardiovaskular.

Tabel 4. 10 Hasil Operator Performance

Accuracy: 71.16%			
	true 0	true 1	class precision
pred. 0	5311	2344	69.38%
pred. 1	1693	4652	73.32%
class recall	75.83%	66.50%	

Hasil evaluasi kinerja model klasifikasi dengan metrik accuracy, precision, dan recall untuk dua kelas, yaitu kelas 0 dan kelas 1. Model memiliki accuracy sebesar 71.16%.

Tabel 4. 11 Correlation Matrix

Attributes	index	id	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio
index	1	-0.003	0.003	0.004	-0.003	-0.002	0.003	-0.003	0.006	0.002	-0.004	0.001	0.004	0.004
id	1.000	1	0.003	0.003	-0.003	-0.001	0.003	-0.002	0.006	0.002	-0.003	0.001	0.003	0.004
age	0.003	0.003	1	-0.022	0.081	0.053	0.026	0.017	0.154	0.098	-0.047	-0.029	0.009	0.238
gender	0.004	0.003	-0.022	1	0.499	0.155	0.006	0.015	-0.035	0.020	0.338	0.170	0.005	0.008
height	0.003	0.003	0.081	0.499	1	0.290	0.005	0.006	-0.050	0.018	0.187	0.094	0.006	0.011
weight	0.002	0.002	0.053	0.155	0.290	1	0.030	0.034	0.141	0.106	0.067	0.067	0.016	0.182
ap_hi	0.003	0.003	0.026	0.006	0.005	0.030	1	0.016	0.023	0.011	0.022	0.001	0.013	0.054
ap_lo	0.003	0.003	0.017	0.015	0.006	0.034	0.016	1	0.024	0.016	0.025	0.010	0.004	0.066
cholesterol	0.006	0.006	0.154	0.035	0.050	0.141	0.023	0.024	1	0.454	0.010	0.035	0.009	0.221
gluc	0.002	0.002	0.098	0.020	0.018	0.106	0.011	0.016	0.452	1	0.004	0.011	0.006	0.089
smoke	0.004	0.003	0.047	0.338	0.187	0.067	0.022	0.025	0.010	0.004	1	0.340	0.025	0.015
alco	0.001	0.001	0.029	0.170	0.094	0.067	0.010	0.010	0.035	0.011	0.340	1	0.025	0.017
active	0.004	0.003	0.009	0.005	0.006	0.016	0.013	0.004	0.009	0.006	0.025	0.025	1	0.036
cardio	0.004	0.004	0.238	0.008	0.011	0.182	0.054	0.066	0.221	0.089	0.015	0.017	0.036	1

Berdasarkan correlation matrix pada Tabel 4.11 atribut *age*, *cholesterol*, *weight*, *dan cardio* dipilih untuk analisis penyakit jantung karena memiliki korelasi yang signifikan terhadap atribut cardio yang dijadikan sebuah label. Atribut *age* (0.238) menunjukkan bahwa semakin bertambah usia, risiko penyakit jantung meningkat, sementara *cholesterol* (0.221) berkaitan dengan kadar kolesterol tinggi yang sering dikaitkan dengan penyakit kardiovaskular. Meskipun nilai korelasi weight (0.182) lebih kecil, ada hubungan positif antara berat badan dan risiko penyakit jantung, khususnya terkait obesitas. Atribut lain seperti *gender*, *height*, *smoke*, *alco*, dan *active* tidak dipilih karena memiliki korelasi rendah atau mendekati nol terhadap *cardio*, sehingga kurang berkontribusi dalam proses analisis.

4.6 Interpretation / Evaluation

Pada tahap interpretasion melakukan ekperimen terhadap nilai K terbaik dengan beberapa tahap yang dilakukan yaitu dengan menguji nilai K dari 2 – 50 seperti di Tabel 4.12.

Tabel 4. 12 Hasil mencari Nilai K

Measure Type	Nilai K	Akurasi Ratio 70:30	Akurasi Ratio 80:20
Mixed Measure	2	61,84%	63,84%
	3	65,45%	66,36%
	4	65,62%	66,69%
	5	67,66%	68,11%
	6	67,28%	68,39%
	7	68,31%	69,01%

	8	68,52%	69,26%
	9	69,00%	69,19%
	10	69,26%	69,57%
	11	69,58%	69,79%
	12	69,54%	69,94%
	13	69,84%	70,09%
	14	69,90%	70,18%
	15	69,83%	70,38%
	16	70,20%	70,34%
	17	70,10%	70,59%
	18	70,30%	70,56%
	19	70,40%	70,75%
	20	70,40%	70,69%
	21	70,32%	70,81%
	22	70,41%	70,83%
	23	70,35%	70,81%
	24	70,61%	70,91%
	25	70,43%	71,16%
	26	70,73%	71,09%
	27	70,50%	70,97%
	28	70,76%	71,01%
	29	70,88%	71,04%
	30	70,80%	71,11%
	31	70,78%	71,06%
	32	70,75%	71,15%
	33	70,78%	71,05%
	34	70,71%	71,03%
	35	70,74%	70,91%
	36	70,77%	70,91%
	37	70,95%	70,74%
	38	70,84%	70,85%
	39	70,93%	70,99%
	40	71,00%	71,05%
	41	70,91%	71,15%
	42	70,88%	71,04%
	43	70,84%	70,94%
	44	70,87%	70,99%
	45	70,81%	70,94%
	46	70,85%	70,91%
	47	70,74%	70,81%
	48	70,75%	70,91%
	49	70,70%	70,94%
	50	70,70%	70,89%

Hasil eksperimen untuk mencari nilai K terbaik dalam algoritma *K-Nearest Neighbors (K-NN)* pada dataset prediksi risiko penyakit kardiovaskular. Dalam tabel 4.12 dua rasio pembagian data diuji 70:30 dan 80:20, yang mengacu pada persentase data yang digunakan untuk pelatihan dan pengujian. Hasilnya menunjukkan bahwa akurasi [22] terbaik untuk rasio 70:30 tercapai pada nilai K 40 dengan akurasi 71,00%, sedangkan untuk rasio 80:20, akurasi terbaik tercapai pada nilai K 25 dengan akurasi 71,16%. Hal ini mengindikasikan bahwa semakin sesuai pembagian datanya, semakin stabil dan akurat model dalam memprediksi risiko penyakit kardiovaskular.

Tabel 4. 13 Hasil Akurasi Ratio 70:30

Accuracy: 71.00%			
	true 0	*true* 1	*class precision*
pred. 0	8449	3973	68.02%
pred. 1	2057	6521	76.02%
class recall	80.42%	62.14%	

Hasil evaluasi dari model klasifikasi yang memprediksi dua kelas (kelas 0 dan kelas 1). Untuk kelas 0, model memprediksi 8449 sampel dengan benar (*True Negative*), tetapi salah memprediksi 2057 sampel sebagai kelas 1 (*False Positive*). Sedangkan untuk kelas 1, model memprediksi 6521 sampel dengan benar (*True Positive*), namun salah memprediksi 3973 sampel sebagai kelas 0 (*False Negative*). Metrik *Precision* menunjukkan ketepatan model dalam memprediksi setiap kelas: untuk kelas 0, precision-nya adalah 68.02%, artinya 68.02% dari prediksi kelas 0 adalah benar; sedangkan untuk kelas 1, precision-nya adalah 76.02%, yang berarti 76.02% dari prediksi kelas 1 adalah benar. *Recall* mengukur seberapa baik model dalam menemukan setiap kelas: untuk kelas 0, recall-nya adalah 80.42%, artinya model berhasil menemukan 80.42% dari semua sampel kelas 0, sementara untuk kelas 1, *recall*-nya adalah 62.14%, yang menunjukkan model berhasil menemukan 62.14% dari semua sampel kelas 1. Secara keseluruhan, model ini memiliki accuracy sebesar 71.00%. Kemudian selanjutnya hasil pada rasio 80:20 seperti pada Tabel 4.14.

Tabel 4. 14 Hasil Akurasi Ratio 80:20

Accuracy: 71.16%			
	true 0	true 1	class precision
pred. 0	5311	2344	69.38%
pred. 1	1693	4652	73.32%
class recall	75.83%	66.50%	

Hasil evaluasi dari model klasifikasi yang memprediksi dua kelas (kelas 0 dan kelas 1), model memiliki accuracy sebesar 71.16%. Untuk kelas 0, model memprediksi dengan benar 5311 sampel sebagai kelas 0 dan salah memprediksi 2344 sampel sebagai kelas 1. Precision untuk kelas 0 adalah 69.38%, yang artinya 69.38% dari semua prediksi kelas 0 adalah benar-benar kelas 0. Di sisi lain, untuk kelas 1, model memprediksi dengan benar 4652 sampel sebagai kelas 1 dan salah memprediksi 1693 sampel sebagai kelas 0. Precision untuk kelas 1 adalah 73.32%, yang menunjukkan bahwa 73.32% dari semua prediksi kelas 1 adalah benar-benar kelas 1. Mengenai recall, model berhasil menemukan 75.83% dari seluruh sampel yang sebenarnya kelas 0 dan 66.50% dari seluruh sampel yang sebenarnya kelas 1. Dengan kata lain, model lebih baik dalam mendeteksi kelas 0 daripada kelas 1. Secara keseluruhan, meskipun model memiliki akurasi yang cukup baik, ada perbedaan dalam kemampuannya mendeteksi kelas-kelas tersebut, dengan sedikit kecenderungan lebih

baik dalam mengenali kelas 0. Adapun Rumus akurasi yang digunakan yaitu :

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

Sehingga dapat dihitung hasil akurasi tertinggi diperoleh dari ratio 80:20 yaitu seperti berikut ini :

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} + \frac{5311+4652}{5311+4652+1693+2344} + \frac{9.963}{14.000} = 0{,}7116.$$

Evaluasi menggunakan confussion matrix dan classification report menunjukkan akurasi 71,16%.

5. KESIMPULAN

Berdasarkan hasil penelitian dan pembahasan yang telah dilakukan, berikut kesimpulan yang dapat diambil :

a. akurasi Model Prediksi dengan Algoritma *K-Nearest Neighbors (K-NN)*:
   Penelitian ini membuktikan bahwa algoritma K-Nearest Neighbors (K-NN) mampu memberikan performa yang kompetitif dalam memprediksi risiko penyakit kardiovaskular. Dengan penggunaan parameter K yang optimal serta teknik penyeimbangan data, akurasi model dapat ditingkatkan secara signifikan, sehingga mempertegas potensi algoritma ini dalam mendukung diagnosis dini penyakit kardiovaskular.
b. penentuan Nilai K Optimal :
   Pemilihan nilai K optimal menggunakan measure type berbasis mixed measure menghasilkan akurasi tertinggi sebesar 71,16%. Temuan ini mengindikasikan bahwa pemilihan jenis measure yang tepat memiliki dampak signifikan terhadap kinerja algoritma, khususnya pada dataset dengan karakteristik data yang beragam.
c. karakteristik Data yang Berpengaruh:
   Hasil analisis mengungkapkan bahwa atribut utama seperti berat badan, kolesterol, dan usia merupakan faktor yang paling berpengaruh terhadap hasil klasifikasi. Temuan ini sejalan dengan literatur medis yang menyebutkan bahwa faktor-faktor tersebut merupakan indikator utama risiko penyakit kardiovaskular.
d. nilai akurasi terbaik dengan perbandingan 2 ratio :
   Pada tahap eksperimen, hasil menunjukkan bahwa akurasi tertinggi untuk rasio 70:30 diperoleh pada nilai K sebesar 40 dengan akurasi 71,00%, sedangkan untuk rasio 80:20, akurasi terbaik tercapai pada nilai K sebesar 25 dengan akurasi 71,16%. Hal ini menunjukkan bahwa pembagian data yang sesuai berkontribusi pada kestabilan dan akurasi model dalam memprediksi risiko penyakit kardiovaskular.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada pihak-pihak terkait yang telah memberi dukungan terhadap penelitian ini.

DAFTAR PUSTAKA

[1] B. Setiaji and P. A. K. Pramudho, "Pemanfaatan Teknologi Informasi Berbasis Data Dan Jurnal Untuk Rekomendasi Kebijakan Bidang Kesehatan," *Heal. J. Inov. Ris. Ilmu Kesehat.*, vol. 1, no. 3, pp. 166–175, 2022, doi: 10.51878/healthy.v1i3.1649.

[2] A. F. Ariani, D. P. Anggraeni, C. Renatasari, S. Informasi, and I. Komputer, "Optimalisasi Administrasi Basisi Data : Strategi Untuk Manajemen Data Yang Efisien Database Administration Optimization : Strategies For Efficient," no. September, pp. 6–7, 2023.

[3] V. Artanti, M. Faisal, and F. Kurniawan, "Klasifikasi Cardiovascular Diseases Menggunakan Algoritma K-Nearest Neighbors (KNN) Classification of Cardiovascular Diseases using K-Nearest Neighbors (KNN) Algorithm," vol. 23, no. 2, pp. 467–479, 2024.

[4] A. Y. S. Saputra and D. Erwandi, "Identifikasi Faktor Risiko Distress Dan Program Penanggulangan Penyakit Kardiovaskular Di Tempat Kerja," *J. Kesehat. Tambusai*, vol. 4, no. 2, pp. 2056–2066, 2023, doi: 10.31004/jkt.v4i2.15953.

[5] F. Putra, H. F. Tahiyat, R. M. Ihsan, R. Rahmaddeni, and L. Efrizoni, "Penerapan

Algoritma K-Nearest Neighbor Menggunakan Wrapper Sebagai Preprocessing untuk Penentuan Keterangan Berat Badan Manusia," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 4, no. 1, pp. 273–281, 2024, doi: 10.57152/malcom.v4i1.1085.

[6] M. F. Kurnia, S. S. Solihat, G. Windarsih, and D. Usmadi, "IDENTIFIKASI OTOMATIS LIMA JENIS RESAK (Vatica spp.) BERDASARKAN BEBERAPA KARAKTER MORFOLOGI DAUN DAN ALGORITMA PEMBELAJARAN MESIN," *Bul. Kebun Raya*, vol. 26, no. 1, pp. 26–37, 2023, doi: 10.55981/bkr.2023.740.

[7] L. Dewa Satria, D. N. Pratomo, and R. N. S. Amriza, "Perancangan Antarmuka Aplikasi Pelaporan Kegiatan Harian Menggunakan Vue Dengan Geolokasi Real-Time dan Push Notifications," *J. Internet Softw. Eng.*, vol. 4, no. 2, pp. 28–33, 2023, doi: 10.22146/jise.v4i2.8977.

[8] R. Tani Vita, "Perancangan Sistem Informasi Penggajian Karyawan Pada Cv. Tri Multi Jaya Yogyakarta," *J. Sist. Inf. dan Sains Teknol.*, vol. 2, no. 1, 2020.

[9] A. Nugroho, M. I. Khomeini, and R. Heraldi, "Comparison of K-Nearest Neighbor and Support Vector Machine Using Binary Dragonfly Algorithm Optimization," *Sistemasi*, vol. 13, no. 1, p. 39, 2024, doi: 10.32520/stmsi.v13i1.2953.

[10] K. Jenis, H. Berdasarkan, F. Pribadi, P. Dewi, P. Purwono, and S. D. Kurniawan, "Pemanfaatan Teknologi Machine Learning pada," pp. 377–387.

[11] F. Fredilio, J. Rahmad, S. H. Sinurat, D. R. H. Sitompul, D. J. Ziegel, and E. Indra, "Perbandingan Algoritma K-Nearest Neighbors (K-NN) dan Random forest terhadap Penyakit Gagal Jantung," *J. Teknol. Inform. dan Komput.*, vol. 9, no. 1, pp. 471–486, 2023, doi: 10.37012/jtik.v9i1.1432.

[12] A. F. Riany and G. Testiana, "Penerapan Data Mining Untuk Klasifikasi Penyakit Jantung Koroner Menggunakan Algoritma Naïve Bayes," vol. 2, no. 1, pp. 297–305, 2023, doi: 10.35957/mdp-sc.v2i1.4388.

[13] N. Novinaldi, B. Harto, and I. Ismael, "Rancang Bangun Sistem Informasi Pengolahan Data Absensi Pegawai dan Perhitungan Tunjangan Kinerja pada KPU Provinsi Sumatera Barat," *J. Pustaka Data (Pusat Akses Kaji. Database, Anal. Teknol. dan Arsit. Komputer)*, vol. 3, no. 2, pp. 46–50, 2023, doi: 10.55382/jurnalpustakadata.v3i2.652.

[14] Z. Zuriati and N. Qomariyah, "Klasifikasi Penyakit Stroke Menggunakan Algoritma K-Nearest Neighbor (KNN)," *ROUTERS J. Sist. dan Teknol. Inf.*, vol. 1, no. 1, pp. 1–8, 2022, doi: 10.25181/rt.v1i1.2665.

[15] D. P. Sinambela, H. Naparin, M. Zulfadhilah, and N. Hidayah, "Implementasi Algoritma Decision Tree Dan Random Forest Dalam Prediksi Perdarahan Pascasalin," *J. Inf. Dan Teknol.*, vol. 5, no. 3, pp. 58–64, 2023, doi: 10.60083/jidt.v5i3.393.

[16] I. A. Sodik and D. M. K. Nugraheni, "Implementation Cobit 2019 for Evaluation of Health Clinic Information System Governance in Central Java," *J. Tek. Inform.*, vol. 3, no. 6, pp. 1549–1556, 2022, doi: 10.20884/1.jutif.2022.3.6.361.

[17] Y. Kurniasari, "Singular Value Decomposition and Discrete Cosine Transform Application for Landsat Satellite Image Enhancement," *J. Sains Dasar*, vol. 2021, no. 1, pp. 16–23, 2021.

[18] L. Ivania Sidora and N. Hanum Harani, "Sistem Rekomendasi Musik Spotify Menggunakan Knn Dan Algoritma Genetika," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 7, no. 4, pp. 2585–2591, 2024, doi: 10.36040/jati.v7i4.7073.

[19] T. Mahesti, K. D. Hartomo, and S. Y. J. Prasetyo, "Penerapan Algoritma Random Forest Dalam Menganalisa Perubahan Suhu Permukaan Wilayah Kota Salatiga," *J. Media Inform. Budidarma*, vol. 6, no. 4, p. 2074, 2022, doi: 10.30865/mib.v6i4.4603.

[20] A. D. W. M. Sidik, I. Himawan Kusumah, A. Suryana, Edwinanto, M. Artiyasa, and A. Pradiftha Junfithrana, "Gambaran Umum Metode Klasifikasi Data Mining," *Fidel. J. Tek. Elektro*, vol. 2, no. 2, pp. 34–38, 2020, doi: 10.52005/fidelity.v2i2.111.

[21] M. M. Ali, T. Hariyati, M. Y. Pratiwi, and S. Afifah, "Metodologi Penelitian Kuantitatif dan Penerapannya dalam Penelitian," *Educ. Journal.2022*, vol. 2, no. 2, pp. 1–6, 2022.

[22] N. H. Alfajr and S. Defiyanti, "METODE RANDOM FOREST DAN PENERAPAN PRINCIPAL COMPONENT ANALYSIS (PCA)," vol. 12, no. 3, 2024.